Mineração de Dados em Biologia Molecular

Planejamento e
Análise de Experimentos

André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Principais tópicos

- Estimativa do erro
- Partição dos dados
- Reamostragem
- Tipos de erro
- Avaliação do desempenho
- Curvas ROC

Estimativa de erro

- Depende do problema:
  - Classificação: considera taxa de exemplos incorretamente classificados
    - Acurácia
  - Regressão: considera diferença entre valor produzido e valor esperado
  - Agrupamento: diferentes critérios
- Média dos erros obtidos em diferentes execuções de um experimento

Estimativa de Erro de Classificação

- Processo de treinamento é utilizado para seleção do modelo
  - Modelo com a complexidade correta (sem overfitting)
- Após construção do modelo, ele pode ser testado com novos exemplos
  - Evitar modelo otimista
  - Conjunto de teste
    - Estimativa não tendenciosa de erro de generalização
  - Comparação de modelos utiliza desempenho em dados de teste
    - Métodos de amostragem

Métodos de amostragem

- Utilizados para avaliar desempenho de um classificador
  - Hold-out
    - Randon subsampling
  - Cross validation
  - Leave-one-out
  - Bootstrap

Hold-out

- Também conhecido como *split-sample*
- Técnica mais simples para estimativa de erro
- Faz uma única partição da amostra em:
  - Conjunto de treinamento: geralmente 1/2 ou 2/3 dos dados
  - Conjunto paras teste: os dados restantes

Hold-out

- Indicado para grande quantidade de dados (ex.: mais de 1000)
- Pequena quantidade de dados
  - Poucos exemplos são usados no treinamento
  - Modelo pode depender da composição dos conjuntos de treinamento e teste
    - Quanto menor conjunto de treinamento, maior a variância do modelo
    - Quanto menor conjunto de teste, menos confiável a acurácia estimada para ele

Hold-out

- Conjuntos de treinamento e teste não são independentes
  - Classe sub-representada em um conjunto será super-representada no outro
    - E vice-versa
- Aproximação pessimista
- Resultados obtidos podem ser pouco significativos
- Solução: utilizar reamostragem

Métodos de reamostragem

- Utilizam várias partições para os conjuntos de treinamento e teste
  - *Random subsampling*
  - *Cross-validation*
    - *Leave-one-out*
  - *Bootstrap*

Random subsampling

- Diferentes partições treinamento-teste são escolhidas de forma aleatória
  - $D_{Trein} \cap D_{Teste} = \varnothing$
  - Taxa de erro é calculada para cada partição
  - Taxa de erro estimada é a média dos erros para as diferentes partições
- Pode obter uma estimativa de erro mais precisa para o desempenho de um modelo

Cross-validation

- Validação cruzada
- Classe de métodos para estimativa da taxa de erro verdadeira
  - K-fold cross-validation
    - Cada objeto participa o mesmo número de vezes do treinamento
      - E apenas uma vez do teste
    - Estratificado
    - Leave-one-out (K = N)

Leave-one-out

- Sua estimativa de erro é praticamente não tendenciosa
  - Média das estimativas tende a taxa de erro verdadeiro
- Computacionalmente caro
  - Geralmente utilizado para pequenos conjuntos de exemplos
  - 10-fold cross validation aproxima leave-one-out
- Variância tende a ser elevada

5 x 2 Cross-validation

- Conjuntos de treinamento e teste com mesmo tamanho
- Dietterich, 1998

Seja um conjunto de N exemplos
Para i = 1 até 5
Dividir N aleatoriamente em duas metades
Usar metade 1 para treinamento e metade 2 para teste
Usar metade 2 para treinamento e metade 1 para teste

Bootstrap

- Funciona melhor que cross-validation para conjuntos muito pequenos
- Forma mais simples de bootstrap:
  - Ao invés de usar sub-conjuntos dos dados, usar sub-amostras
    - Cada sub-amostra é uma amostra aleatória com substituição do conjunto total de exemplos
    - Cada conjunto de treinamento têm o mesmo número de exemplos do conjunto total
    - Os exemplos que restarem são utilizados para teste

Bootstrap

- Se conjunto original tem N exemplos
  - Amostra de tamanho N tem ≈ 63,2% dos exemplos originais
- Processo é repetido b vezes
  - Resultado final = média dos b experimentos
- Existem diversas variações
  - Como calcular a acurácia do classificador
  - .632 bootstrap

Erro de classificação

- Principal objetivo de um modelo é classificar corretamente para novos exemplos
  - Errar o mínimo possível
    - Minimizar taxa de erro
  - Geralmente não é possível medir com exatidão essa taxa de erro
    - Ela deve ser estimada

Estimativa de erro de classificação

- Acurácia
  - Trata as classes igualmente
  - Pode não ser adequada para dados desbalanceados
    - Classe rara é mais interessante que a majoritária
    - Pode prejudicar desempenho para classe minoritária

Classificação binária

- Dois tipos de erro:
  - Classificação de um exemplo N como P
    - Falso positivo (alarme falso)
      - Ex.: Diagnosticado como doente, mas está saudável
  - Classificação de um exemplo P como N
    - Falso negativo
      - Ex.: Diagnosticado como saudável, mas está doente

Tipos de erro
Classe positiva é, em geral, a classe de maior interesse
Ou com menos exemplos
Em alguns casos, os erros têm igual importância
Em outros, erros diferentes têm conseqüências diferentes
No exemplo anterior, qual é pior (tem mais custo)? falso negativo ou falso positivo?
27/09/2012
André de Carvalho - ICMC/USP
19

Tipos de erro
Matriz de confusão (tabela de contingência) pode ser utilizada para distinguir os erros
Base de várias medidas
Pode ser utilizada com 2 ou mais classes
Classe predita
Classe verdadeira
1 2 3
1 25 0 5
2 10 40 0
3 0 0 20
27/09/2012
André de Carvalho - ICMC/USP
20

Avaliação de desempenho
Matriz de confusão para 200 exemplos divididos em 2 classes
Classe predita
Classe verdadeira
p n
P 70 30
N 40 60
Classe predita
Classe verdadeira
p n
P VP FN
N FP VN
27/09/2012
André de Carvalho - ICMC/USP
21

Avaliação de desempenho
Medidas de desempenho
Taxa de FP (TFP) = FP / (FP + VN)
(Alarmes falsos)
Taxa de FN (TFN) = FN / (VP + FN)
Erro do tipo I
Erro do tipo II
Classe predita
Classe verdadeira
p n
P VP FN
N FP VN
27/09/2012
André de Carvalho - ICMC/USP
22

Avaliação de desempenho
Medidas de desempenho
Taxa de FP (TFP) = FP / (FP + VN)
(Alarmes falsos)
Taxa de VP (TVP) = VP / (VP + FN)
Custo
Benefício
Classe predita
Classe verdadeira
p n
P VP FN
N FP VN
27/09/2012
André de Carvalho - ICMC/USP
23

VP / (VP + FN)
FP / (FP + VN)
Exemplo
Avaliação de 3 classificadores
Classe predita
Classe verdadeira
p n
P 20 30
N 15 35
P 70 30
N 50 50
P 60 40
N 20 80
Classificador 1
TVP =
TFP =
Classificador 2
TVP =
TFP =
Classificador 3
TVP =
TFP =
27/09/2012
André de Carvalho - ICMC/USP
24

$\frac{VP}{VP+FN}$ $\frac{FP}{FP+VN}$
Exemplo
Avaliação de 3 classificadores
Classe predita
p n
Classe verdadeira
P 20 30
N 15 35
Classificador 1
TVP = 0.4
TFP = 0.3
Classe predita
p n
Classe verdadeira
P 70 30
N 50 50
Classificador 2
TVP = 0.7
TFP = 0.5
Classe predita
p n
Classe verdadeira
P 60 40
N 20 80
Classificador 3
TVP = 0.6
TFP = 0.2
27/09/2012
André de Carvalho - ICMC/USP
25

Avaliação de desempenho
Medidas frequentemente utilizadas
TFP = $\frac{FP}{VN+FP}$
(Erro tipo I)
TFN = $\frac{FN}{VP+FN}$
(Erro tipo II)
Precisão = $\frac{VP}{VP+FP}$
TVP = $\frac{VP}{VP+FN}$
Sensibilidade
Revocação (Recall)
Especificidade = $\frac{VN}{VN+FP}$ = 1–TFP
Acurácia = $\frac{VP+VN}{VP+VN+FP+FN}$
Medida-F1 = $\frac{2}{1/prec+1/rev}$
27/09/2012
André de Carvalho - ICMC/USP
26

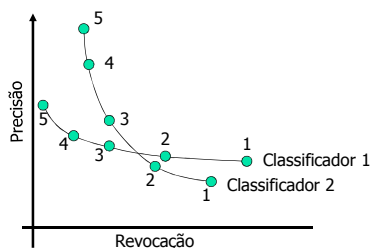
Revocação X Precisão
Revocação (recall) $\frac{VP}{VP+FN}$
Taxa com que classifica como positivos todos os exemplos que são positivos
Nenhum exemplo positivo é deixado de fora
Precisão $\frac{VP}{VP+FP}$
Taxa com que todos os exemplos classificados como positivos são realmente positivos
Nenhum exemplo negativo é incluído
27/09/2012
André de Carvalho - ICMC/USP
27

Sensibilidade X Especificidade
Sensibilidade $\frac{VP}{VP+FN}$
Taxa com que os exemplos positivos são classificados como positivos
Igual a revocação
Especificidade $\frac{VN}{VN+FP}$
Taxa com que exemplos negativos são classificados como negativos
Nenhum exemplo negativo é deixado de fora
27/09/2012
André de Carvalho - ICMC/USP
28

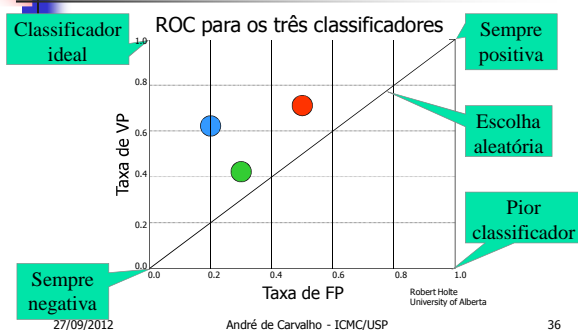
Avaliação de desempenho
Medida-F
Média harmônica ponderada da precisão e da revocação
$\frac{(1+\alpha)\times(prec\times rev)}{\alpha\times prec+rev}$
Medida-F1
Precisão e revocação têm o mesmo peso
$\frac{2\times(prec\times rev)}{prec+rev} = \frac{2}{1/prec+1/rev}$
27/09/2012
André de Carvalho - ICMC/USP
29

Exemplo
Seja um classificador com a seguinte matriz de confusão, definir:
Acurácia
Precisão
Revocação
Especificidade
Classe predita
p n
Classe verdadeira
P 70 30
N 40 60
27/09/2012
André de Carvalho - ICMC/USP
30

Exemplo
Acurácia = $\frac{VP + VN}{VP + VN + FP + FN}$
Precisão = $\frac{VP}{VP + FP}$
Revocação = $\frac{VP}{VP + FN}$
Especificidade = $\frac{VN}{VN + FP}$
Predito
p n
P VP FN
N FP VN
Verdadeiro
p n
P 70 30
N 40 60
27/09/2012
André de Carvalho - ICMC/USP
31

Exemplo
Acurácia = $\frac{VP + VN}{VP + VN + FP + FN}$ = (70 + 60) / (70 + 30 + 40 + 60) = 0.65
Precisão = $\frac{VP}{VP + FP}$ = 70/(70+40) = 0.64
Revocação = $\frac{VP}{VP + FN}$ = 7/(70+30) = 0.70
Especificidade = $\frac{VN}{VN + FP}$ = 60/(40+60) = 0.60
Predito
p n
P VP FN
N FP VN
Verdadeiro
p n
P 70 30
N 40 60
27/09/2012
André de Carvalho - ICMC/USP
32

Observação
Precisão
Revocação
Classificador 1
Classificador 2
5
4
3
2
1
27/09/2012
André de Carvalho - ICMC/USP
33

Gráficos ROC
Do inglês, *Receiver operating characteristics*
Medida de desempenho originária da área de processamento de sinais
Muito utilizada nas áreas médica e biológica
Mostra relação entre custo (TFP) e benefício (TVP)
27/09/2012
André de Carvalho - ICMC/USP
34

Exemplo
Colocar no gráfico ROC os 3 classificadores do exemplo anterior
Classificador 1
TFP = 0.3
TVP = 0.4
Classificador2
TFP = 0.5
TVP = 0.7
Classificador 3
TFP = 0.2
TVP = 0.6
27/09/2012
André de Carvalho - ICMC/USP
35

Gráficos ROC
ROC para os três classificadores
Classificador ideal
Sempre positiva
Escolha aleatória
Pior classificador
Sempre negativa
Taxa de VP
Taxa de FP
Robert Holte
University of Alberta
27/09/2012
André de Carvalho - ICMC/USP
36

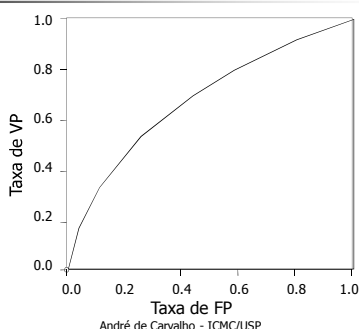
Gráficos ROC

- Classificadores discretos produzem um simples ponto no gráfico ROC
  - ADs e conjuntos de regras
- Outros classificadores produzem uma probabilidade ou escore
  - RNAs e NB
- Curvas ROC permitem uma melhor comparação de classificadores
  - São insensíveis a mudanças na distribuição das classes

Curvas ROC

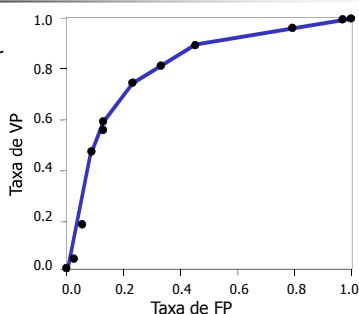
- Mostram ROC para diferentes variações
- Classificadores que geram escores ou probabilidades
  - Diferentes valores de *threshold* podem ser utilizados para gerar vários pontos
    - Cada valor de *threshold* produz um ponto diferente
    - Ligação dos pontos gera uma curva ROC

Curvas ROC

Classificador Escore/ Probabilístico

Curvas ROC

- Classificadores que geram valores discretos
  - Podem ser convertidos internamente para gerar escores
    - Para ADs, diferentes thresholds para números de exemplos positivos que tornam a classe positiva
  - Podem ser combinados em comitês
    - Threshold para votos dos classificadores individuais forma escore

Curvas ROC

Classificador Discreto

Área sob a curva ROC (AUC)

- Fornece uma estimativa do desempenho de classificadores
- Gera um valor continuo no intervalo [0, 1]
  - Quanto maior melhor
  - Adição de áreas de sucessivos trapezóides
- Um classificador com maior AUC pode apresentar AUC pior em trechos da curva
- É mais confiável utilizar médias de AUCs

Área Sob Curvas ROC
Nenhuma Discriminação
Taxa de VP
1.0
0.8
0.6
0.4
0.2
0.0
Área = 0.5
0.0 0.2 0.4 0.6 0.8 1.0
Taxa de FP
27/09/2012
André de Carvalho - ICMC/USP
43

Área Sob Curvas ROC
Discriminação Perfeita
Taxa de VP
1.0
0.8
0.6
0.4
0.2
0.0
Área = 1.0
0.0 0.2 0.4 0.6 0.8 1.0
Taxa de FP
27/09/2012
André de Carvalho - ICMC/USP
44

Área Sob Curvas ROC
Taxa de VP
1.0
0.8
0.6
0.4
0.2
0.0
Área = 0.66
0.0 0.2 0.4 0.6 0.8 1.0
Taxa de FP
27/09/2012
André de Carvalho - ICMC/USP
45

Área Sob Curvas ROC
Taxa de VP
1.0
0.8
0.6
0.4
0.2
0.0
Área = 0.82
Área = 0.66
0.0 0.2 0.4 0.6 0.8 1.0
Taxa de FP
27/09/2012
André de Carvalho - ICMC/USP
46

Avaliação de Desempenho
Teste de Hipóteses
Compara dois desempenhos
Teste t
Teste McNemar
Teste t pareado de 5x2 CV
Compara vários desempenhos
Teste Feelders e Verkooijen
Teste de Friedman
ANOVA
27/09/2012
André de Carvalho - ICMC/USP
47

Considerações Finais
Estimativa do erro
Avaliação do desempenho
Erro
Tempo de resposta
Memória
Representação
Medidas
Gráficos e curvas ROC
Teste de hipóteses
27/09/2012
André de Carvalho - ICMC/USP
48

Perguntas
27/09/2012
André de Carvalho - ICMC/USP
49